Ishan Madhuranga

Dublin, Ireland

+353 830 242 827 | contact@excellenttorch.com | LinkedIn | Portfolio

Professional Profile

AI/Machine Learning Engineer with 4+ years of experience in delivering end-to-end AI solutions. Skilled in **Python**, **C++**, and **AWS/Azure** Cloud Deployment, with a track record of developing optimized AI models, AI behaviour-based gameplay systems. Strong background in Generative Models (**GAN**s), **Transformer** Models, **LLM**s, **LangChain** + **RAG** pipelines, and scalable **inference engines**, delivering impactful results across Industry, Academic, and Freelance Projects.

Education

MSc (Hons) Artificial Intelligence — **Jan. 2023 – Jan. 2025**
Dublin Business School — *Dublin, Ireland*

BSc (Hons) Cyber Security Management — **April. 2021 – April. 2022**
Solent University — *Southampton, UK*

Skills

- **Languages:** Python, C++, C#, HTML/CSS, JavaScript/TypeScript
- **Frameworks:** PyTorch, Tensorflow, OnnxRuntime, FastAPI, Flask, React, React Three Fiber, Node.js, FastAPI, LangChain, LlamaIndex, MatLab
- **AI/ML:** Deep Learning, NLP, Computer Vision, Reinforcement Learning, Transformer Models, Generative AI, and Model Optimization.
- **Tools:** Visual Studio, AWS, Azure DevOps, Git, Docker, Kubernetes
- **Databases:** MySQL, MongoDB, PostgreSQL, Amazon RDS
- **Methodologies:** Agile, SDLC

Experience

Sleepless Inc. — **Sept. 2024 – Jun. 2025**
Lead AI Gameplay Programmer — *Dublin, Ireland*

- Developed and optimized **NPC AI** behaviours through Blueprints and **C++** using **Unreal Engine 5**.
- Improved game efficiency by profiling performance bottlenecks, optimizing AI pathfinding, and refining blueprint execution with C++, resulting in a 60% reduction in runtime.
- Collaborated in Agile sprints and followed SDLC practices to iteratively develop through **GIT**/**Diversion**.

British College of Applied Studies — **Nov. 2021 – Apr. 2023**
Artificial Intelligence Engineer — *Colombo, Sri Lanka*

- Contributed to academic research projects involving **GAN**s, autonomous systems, and reinforcement learning, improving model performance overall.
- Built and deployed **RESTful** APIs using **FastAPI** and **Flask**, reducing system runtime.
- Deployed AI solutions on **AWS** and **Azure**, ensuring scalable experimentation and delivery.
- Documented all development processes, including model training, evaluation, and deployment strategies, in line with academic and industry standards.

Freelance AI/ML Specialist — **Aug. 2019 – Feb. 2025**
Fiverr, Upwork & Direct Contracts — *(Remote) Sri Lanka*

- Delivered AI and ML solutions to clients, leveraging **Python** to build and deploy models for deep learning, computer vision, and data analytics tasks.
- Conducted model optimization and integration for real-time applications.
- Developed end-to-end technical solutions for data pre-processing and model deployment using **Amazon ECS**/**Azure Kubernetes** and tailored every minor adjustment for client satisfaction.

Projects

Fast Neural Style Transfer In a 3D Environment in Real Time | Python, PyTorch, ONNX

- Developed a custom **Transformer** model for a real-time neural style transfer system for 3D environments.
- Implemented intelligent NPC behaviour using **ConvAI** for conversations with a custom fine-tuned **BART LLM** model with a **RAG** pipeline running on the backend to give unique personalities.
- Achieved seamless rendering while maintaining visual quality with Custom Depth Filters in **Unreal Engine 5**.
- Quantized the ONNX Models for better performance.

AI Research Assistant with RAG & LLM Integration | Python, React, TypeScript

- Built an end-to-end AI solution with a RAG pipeline (**LangChain** + **FAISS**) and integrated LLMs for multi-turn reasoning, summarization, and document analysis.
- Developed and deployed a **FastAPI** backend with a React + TypeScript dashboard, enabling real-time, client-ready interactions.

Optimized Inference Engine for ARM/Mobile Devices | C++, ONNX, TensorRT

- Optimized deep learning models for real-time performance on ARM/Mobile devices by converting **PyTorch** models to ONNX and accelerating inference with TensorRT.
- Applied quantization and pruning techniques to shrink model size and cut latency, achieving faster and more efficient execution on **NVIDIA Jetson Nano** hardware.
- Packaged the solution in Docker with automated tests and monitoring tools, ensuring reliability and scalability for production like deployments.

Autonomous Self-Driving Car (Prototype) | Python, NVIDIA Jetson Nano

- Implemented an autonomous self- driving vehicle using a custom-trained **CNN** model deployed on a single board computer.
- Real-time inference from a single **RGB** camera for dynamic navigation decisions.
- Applied **Threading** for monitoring driver fatigue using **HOG**-based face detection and facial landmark analysis.
- Implemented Real time Motor RPM Adjustments based on Inference Results.